The Report Committee for Maria Renatovna Doulatova

Certifies that this is the approved version of the following report:

Self-Attributions and Other-Attributions

Revisited From a Neural Perspective

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Daniel A Bonevac

Co-Supervisor:

Jesse J. Prinz

Adam Pautz

Rick Grush

Self-Attributions and Other-Attributions

Revisited From a Neural Perspective

by

Maria Renatovna Doulatova, B.A.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

December 2012

Dedication

For my parents, Olga and Renat.

Acknowledgements

I thank Tyler Burge, Dan Bonevac, Rick Grush, Jesse Prinz, Sinan Dogramaci, Adam and Anna Pautz for valuable contributions to my thinking on this topic.

Abstract

Self-Attributions and Other-Attributions

Revisited From a Neural Perspective

Maria Renatovna Doulatova, M.A.

The University of Texas at Austin, 2012

Supervisors: Dan Bonevac, Jess Prinz

Caruthers argues that the mindreading capacity and the introspective capacity are in fact one and the same capacity. This single capacity relies on the same sub-personal "interpretive" mechanism that takes sensory information as input and produces attitudes as output. I use neuroscience research to show that if the “interpretive mechanism” exists, and moreover that it operates in accordance to Caruthers’ description in mindreading tasks, (e.g. detecting external cues and paying attention to others’ behavior), then this operation would have to be handled or implemented at the neural level by the Task Oriented Neural Network. On the other hand, it is well known that self-referential thought, including introspective thought is handled by the Default Mode Network. This consequence is problematic for the view that self and other attitude attributions are done by the *same* mechanism. The same cognitive operation can not be implemented by two distinct neural networks that are in competition with one another. Moreover, the Default

Mode neural network and the Task Oriented networks implement such different types of thinking that they oppose and interrupt one another's functioning. If the only difference between the two networks were that one simply handles a larger *quantity* of information than the other, then they wouldn't be in competition. It appears that there is indeed something special about the very nature of self-referential information such that it determines the *type* of operations involved in its processing.

Table of Contents

INTRODUCTION

Caruthers argues that the mindreading capacity (attributing judgments to others e.g. "she does not think pasta is a good meal") and the introspective capacity (attributing judgments to self e.g. "I do not think pasta is a good meal") are in fact one and the same capacity (Caruthers, 2010). This single capacity relies on the same sub-personal "interpretive" mechanism that takes sensory information as input and produces attitudes as output. According to Caruthers, the only difference between the self attributions and the other attributions is quantitative and not qualitative in kind. That is, there is simply more sensory evidence available for self attributions than for other attributions. Importantly the type of access to one's attitudes and others' attitudes is the same. Intuitions of authority and phenomenological immediacy (when it comes to self attributions) are explained away as simply due to practice and plethora of sensory information. Confabulation experiments have shown infallibility of self attributions to be empirically lacking. That is, while we think we know our own judgments, experiments show that we often confabulate and make stuff up. Moreover, Caruthers argues that there is nothing unique about phenomenological immediacy since we often make attributions to others' judgments with similar quickness and ease.

This paper is aimed at challenging the above claims. Namely I will *contest* the following subparts of the above idea.

(1) The attitudes of the self and of the other are always interpreted using the same kind of mechanism.

(2) Evolution favors the existence of a single "other-directed" mindreading mechanism (with indirect introspective ability being a mere by-product of such pressures)

This paper will contain two chapters. In chapter one, I will lay out Caruthers' view as presented in his paper "Introspection: Divided and Partly Eliminated" and elaborated on in his book *The Opacity of Mind*. In chapter two, I will present an objection to his view.

A preliminary consideration is in order. While Caruthers' main target is the idea that we have direct introspective access to our judgments and attitudes, I will remain neutral on this claim throughout this paper. I will mainly concern myself with taking apart the argument he makes against direct introspective access. These challenges will be grounded in Caruthers' favored playing field: empirical evidence. Specifically I will illustrate how current neuroscience research cannot accommodate any version of a "self/other parity" account. Philosophers who defend this (neurological) account argue that the process by which we come to know our own attitudes is the same type of process by which we come to know the attitudes of others. I will show that it can not be the same process since two distinct neural networks handle each process respectively. Moreover, these two networks are in competition with one another.

The upshot of using this neuroscience research is the following. If the "interpretive mechanism" exists, and moreover that it operates in accordance to Caruthers' description in mind reading tasks, (e.g. detecting external cues and paying attention to others' behavior), then this operation would have to be handled or implemented at the neural level by the Task Oriented Neural Network. On the other hand, it is well known that self-referential thought, including introspective thought is handled

by the Default Mode Network. This consequence is problematic for the view that self and other attitude attributions are done by the *same* mechanism. The same cognitive operation can not be implemented by two distinct neural networks that are in competition with one another.

Furthermore, I will offer an alternative evolutionary story from that of Caruthers'. He conjectures that an organism' fitness would be better served if he were to rely on a single "do it all" interpretive mechanism for self and other attributions. Why waste the extra energy developing two mechanisms when one serves both functions? I will explain how an organism's fitness would be increased if there really were two distinct cognitive mechanisms for engaging in self-referential and non-self-referential thought.

In sum, these considerations will show that Caruthers' thesis that same sub-personal "interpretive" mechanism is responsible for self and other attributions is really not the *best explanation* for all the relevant empirical findings.

Chapter 1. Challenge to Introspection: ISA Theory

1.1 WHAT IS "INTROSPECTION"?

In his paper *Introspection: Divided and Partly Eliminated*, Caruthers argues that there is no such thing as introspective access to one's judgments (*Caruthers defines judgments as events of belief-formation, and decision as events of intention formation), attitudes, and decisions.

The type of access in question is a way of learning about your own (fairly current) mental states that is not available to others. Specifically, Caruthers defines introspection such that introspective judgments do not (only) appeal to the subject's behavior or circumstances as inputs (p.77). In other words, if introspective access were available to you, you could exercise it in the following way. If someone were to ask you if you agree that, "Russian gymnasts are the best in the world," all you would have to do is "focus inwards" and you would, via introspection, see whether you in fact hold this attitude. In order to make this view somehow more plausible at first blush, it is necessary to note that Caruthers does indeed grant introspective access perceptual experiences, imagery and inner speech. Going back to our example, if you are sitting and the following image springs to mind "Russian gymnasts all adorned in gold medals with a gold-like halo shinning around their heads", this image will serve as a pretty vital piece of input to your belief formation.

In order to see what species of views Caruthers is challenging with his thesis, it would be useful to go over several vital features typically attributed to introspective

access in the literature. Introspective access is always described with **two** characteristic features: not available to others (i.e. special or peculiar), and authoritative.

The type of knowledge that results from real introspective access is supposed to be particularly secure. That is, you are the authority on whether you hold this attitude or not. It would be ludicrous for someone to come up to you and inform you, to your surprise, that you in fact do not hold this attitude. All someone could do in protest is point to some of your external behavior he came to observe. For example, he could say "I don't think you really judge that Russian gymnasts are the best in the world, since you were booing them during the Olympics". This observation of your behavior is of course not a foolproof or "reliable" indicator of your true attitudes. In reply you could say "Yes, I was making booing sounds but that's because I wanted to appear as if I don't hold that they are the best in the world". There is nothing really that your friend could say in response. In other words, you have the final say on what is going on "inside your head". So while other people can observe your behavior and theorize about its relation to your attitudes, all you have to do is "introspect" and your attitudes will become inwardly apparent to you. Moreover, while you can be mistaken about lots of things i.e. the answer to your logic homework, your true national origins, the size of your pen, or even the color of your room; you cannot be mistaken about what attitudes you hold or what judgments you make. That is your reasoning could be off in performing modus tollens, your parents could have been lying to you and you are in fact adopted, the stripes on your pen could be making it look wider than it actually is, you could be experiencing some strange hallucination and your walls could appear purple to you. However, if you focus inwards, you will be able indubitably tell what your attitudes are. Since there is no room for error,

the introspective access is generally held to be "direct" (It is notable that *authority* can come in milder varieties).

In summary, it would be useful to go over a few conditions that mark a truly introspective process. The "mentality condition" is simply the condition that delineates the targets of an introspective process. According to this condition, in order for a process to count as introspective, it provides access to your own mental events, and not to events that are going on outside your window. The "first-person condition" states that in order to count as introspective, a process must be generating knowledge only about your own mental states. Simply staring intensely at yourself in the mirror in an attempt to determine your mental goings on does not amount to introspection per se, since you can just as easily stare at other people. Again, you have to somehow "focus inwards" at your own mental goings on in order to count as introspecting them. According to this condition, you can not introspect the mental states of others. The "temporal proximity condition" dictates that introspection could only be attempted as a way of getting on your fairly current mental goings on. If you try to get onto what you thought about the Russian gymnastics team back in the Soviet era, you would no longer be introspecting, but merely remembering your mental states. The next condition is particularly mysterious and yet has been taken to be obvious by philosophers and folk psychologists alike. The "directness condition" states that you get onto your mental states "directly". That is, you don't have to pull out your pocket dictionary and consult your latest Facebook status updates in order to know what your judgments currently are. You simply "look within" and see. Notably, most philosophers do not hold that no "computations" are made when an attitude is introspected. Some grant that the mind has an architecture that is complex

and perhaps performs some sub-personal computations before a full fledged thought is formed (Lycan, 1987, 1996). What is important is that these computations aren't inferences that take the subject's own behavior and circumstances as premises.

Another rather mysterious condition is called the "detection condition". This condition implies that when you introspect, you get onto something that is already formed. In this way, introspection is similar to fishing (with a motion detector). Your inner attitudes and judgments are fish in the pond and when you catch a fish you bring your attitudes to light. That is, while you might not have been paying attention to some of your attitudes while watching TV, when you focus within and "detect" the attitude, you inevitable become aware of it. This notion is further elaborated upon in the "effort condition", which states that an introspective process is one that requires some effort. To continue with the fishing analogy, fish don't just come up to you; you have to throw a line into the water and catch it. In other words, you have to "focus within" and you will get onto your attitudes. These attitudes are already there, so to speak, waiting to be detected. In your effort to introspect, you don't just proceed to make attitudes up as you go, they are already formed.

This paper will center around Caruthers' attack on the "first person" condition. This condition is upheld by *the Transparency Theory of Self-Knowledge.* The transparency theory aims to do the following. Firstly, it offers a concrete and clear proposal of a non-extravagant mechanism for acquiring self-knowledge. Secondly, it purports to vindicate authoritative and special access. Byrne offers a particularly clear exposition of this approach. He defends a transparency *method*, or *epistemic rule*:

BEL: if *p*, then believe that you believe that *p*.

According for Byrne, following the epistemic rule requires *recognizing* that its antecedent obtains, and then doing what its consequent prescribes *because of* such recognition.

This method is supposed to secure the subject authoritative access described above. According to Byrne, if you merely try to follow the rule, the prescribed belief will be true. Trying to follow BEL or any other epistemic rule involves *believing* that its antecedent obtains (*p* is true), but being wrong about this (the antecedent does not obtain; *p* is not true), and thus failing to *know* or to *recognize* that the antecedent obtains (failing to know *p*).

The upshot is that the subject believes-but-doesn't-know *p*, which of course suffices to make true the item of self-knowledge that the rule instructs the subject to have: believe that you believe *p*. Since the method suffices to make the output item of self-knowledge true, i.e. provides a kind of reliable basis for true belief, the output item qualifies as *knowledge*.

Moreover, according to Byrne, the transparency method secures special access. Byrne suggests the rule could in principle be applied to other people but would not retain. This method is non-extravagant. If we can reason, that is, if we can follow epistemic rules, then we can follow this one to acquire self-knowledge.

Caruthers' account of introspection is a type of "self/other parity" account. That is he denies 1st person access, as well as authoritative and special access. Contra Byrne, he argues that the process by which we come to know our own attitudes is the same type of process by which we come to know the attitudes of others.

1.2 CARUTHERS' MODEL OF INTROSPECTION

Caruthers defends an "Interpretive Sensory Access" theory of self knowledge (*Opacity of Mind*, preface, xii). This theory goes against the 1st person condition and states that our mode of access to our own thinking is the same as our mode of access to the thinking of other people (*ibid).* According to Caruthers, judgments are current evens of belief-formation and decisions are "events that create novel activated intentions" (*ibid*)

Here are the relevant predictions made by the ISA theory:

(a) There is a single mental faculty underlying our attributions of propositional attitudes, whether to ourselves or to others.

(b) The mental faculty in question evolved to sustain and facilitate outward looking or rather other-directed, forms of cognition.

Caruthers' argument is a type of Inference to the Best Explanation. He appeals to research in cognitive science to propose a model of how we come to have access to our judgments. Roughly speaking, this model of cognition is the following: "the human mind exemplifies a perception/belief/desire-making architecture (pp. 79-80). According to this model, perceptual outputs are "globally broadcast" to a wide range of concept-using systems in the mind.

This model also explains psychological findings on confabulation. He takes the relative acceptance of this model among cognitive scientists to be indicative of its validity and its explanatory power for mental events like introspection[1]. Let us look at this model and his interpretation of it more closely.

[1] This model is found to be highly suspect by neuroscientists (Poldrack, personal correspondence)

> On this account, there are a range of perceptual systems (visual, auditory, somatosensory, etc.) which broadcast their outputs to a set of conceptual systems. Some of these generate judgments, some create new goals, and some generate decisions and intentions for action. Each of these conceptual systems can store its outputs in memory, and can access and activate those stored representations when reasoning. Included among the systems for generating judgments and beliefs is a **mindreading faculty**, which produces higher-order judgments about the mental states of others and of oneself.

The central feature of this architecture for our purposes is what Caruthers refers to as the "mindreading faculty". This is the faculty, or a type of mechanism, that receives input in the form of sensory information (i.e. perceptual, somatosensory, and olfactory). According to Baars, (Baars, 2003) some of this sensory information is already "infused" with concepts that serve to chunk or categorize it into recognizable bits. The idea is that something with a certain sensory light pattern is "matched" to a conceptual "template" (e.g. coffee cup). All of these sensory subsystems are operating as "slaves" to the working memory or the "workspace" which receives "broadcasts" from the slave subsystems. Other faculties "look onto" the workspace and pick information that is relevant to their operations. Caruthers considers the "mindreading faculty" to be one of such onlookers or "consumers" or workspace "broadcasts".

This picture is somewhat analogous to the Facebook architecture. The workspace is sort of like your "wall" where other subsystems (your friends or apps) "broadcast" on. Your "mindreading" faculty is in turn sort of like your very politically minded self who scans your wall for any posts relevant to politics. It disregards all posts that are not relevant to politics. It then complies and combines the political posts in a meaningful way and comes up with a report on the current political atmosphere of your wall. Your "politically minded self" can also scan the walls of your friends and do the same

procedure with their walls. All it has to do is look for politically oriented posts, apply some sort of interpreting mechanism (e.g. some theory of how posts of this nature come together to reflect the overall political atmosphere of a wall) and spit out a pronouncement on the political atmosphere of a wall (be it your own wall or the wall of your friends). Let us return to an example introduced in the beginning of the paper. Say you feel as if you introspected the judgment that you think that the Russian gymnastics team is the best in the world. According to Caruthers, this introspective pronouncement comes about in the following way. The sensory subsystems have broadcasted the following input to the mindreading faculty. You feel exhilarated and excited when the Russian gymnasts earn a perfect score on a certain apparatus. You find yourself jumping up and down when their score is announced. You feel upset when the judges take deductions from their score. You have cleared your entire day's schedule just to be able to watch them compete. You take bathroom breaks only when the gymnasts of other countries are performing. The mindreading faculty receives these broadcasts and applies the following theory to the data. I never cheer for any team except the Russian team. Anyone who cheers for a team must really like the team. I never deny myself snack breaks when a sports event is televised, and freely take breaks when any team performs except when the Russians perform. Anyone who makes such sacrifices must really like the team… . It is easy to see how the rest of the "interpretation" might run.

According to Caruthers, the same type of interpretive mindreading process takes place when you discern that your friend holds that the Russian gymnastics team is the best in the world. You see her cheering for the Russian team, taking breaks only when the other teams are performing and never when the Russian team is performing, yelling at the

TV when the judges deduce points from the Russian team and cheering on when deductions are made to other teams and so on. Your interpretive mechanism takes all this evidence into consideration, applies a theory "anyone who acts this way must really like the team" and attributes an attitude to your friend "you must really hold that the Russian team is the best gymnastics team in the world".

You might naturally protest that you feel that you do not engage in any such interpretive activity. You might add that when you need to figure out whether you hold that the Russian gymnastics team is the best in the world, all you find yourself doing is looking inwards and the attitude simply presents itself to you. You might protest "I do not need to compile evidence and apply some theory to get onto my own attitude." It is ridiculous to think that you look at you own external behavior and think to yourself "wow I am cheering in joy. People generally cheer in joy when they like something. Hence, I must really like this team!"

Caruthers addresses this intuition of immediacy in the following way. According to Baars' model of cognitive architecture, the mindreading mechanism performs its interpretations sub-consciously. That is why you are hardly ever actually aware of making such interpretations and it seems as if these self attributions just present themselves to your consciousness. Consider the Facebook analogy presented above. Imagine that the "politically minded self" is actually not you actively searching various wall posts, trying to fit the data into some theory of politically minded posts and how they come together to produce certain "political atmosphere of the wall" pronouncements. Imagine instead that there is actually just an application that you subscribe to on your profile. It automatically scans your wall, noting various posts with

political content, and automatically uses its program to spit out a given "wall political atmosphere" pronouncement. All you have to do is see its pronouncement and not even consciously note the various posts that went into it or the manner in which they contributed to this particular pronouncement. Furthermore, imagine that this automatic application can scan not only your own wall, but also the walls of all of your friends and even random Facebook members you happen to encounter. Its programming interprets all walls in the same way. All you become aware of is the pronouncement or the attribution. Going back to our previous example, you hardly ever find yourself trying to piece together the behavior or your friend. You don't sit there taking notes on all the instances when she appears particularly excited about the winnings of the Russian team and only smiled a little bit about the winnings of other teams. All you feel is the attribution "she thinks that the Russian gymnastics team is the best!" Caruthers points to this phenomena to show that your intuitions of "immediacy" when it comes to self attributions actually also exist for the attributions you make to others (unless their behavior is somehow particularly puzzling, you never catch yourself consciously trying to figure it out or piece it together).

The only difference is that it has access to many more posts on your wall than on the walls of others. Naturally, the more information a mechanism has, the more accurately it tends to perform. So you see it spiting out pronouncements about the political atmosphere of your wall and somehow they turn out to be fairly reliable. On the other hand, more mistakes creep up when it targets other walls. Going back to our example, when making self attributions about your judgment that the Russian gymnastics team is the best, your mindreading faculty has access to your heart rate, the strength that

you apply to your jumping up and down and so forth. On the other hand, the mindreading mechanism doesn't have direct access to the heart rate of your friend. You can see her jumping up and down but you don't feel the strength that went into the jump. Notably, you also don't have access to her internal monologue or her inner speech.

This model provides a neat explanation for the plethora of confabulation experiments. Notably, the transparency model can offer no explanation for the prevalence of confabulation in normal subjects. Byrne would have to retreat to the view that there has been a breach of rationality in all these cases.

These experiments are set up to show the existence of an interpretation mechanism. Specifically, various cues are presented to the subject's slave systems. These cues are then allegedly broadcast for the interpretation mechanism to consume. These cues are cleverly arranged to ultimately "trick" the interpretation mechanism into spitting out a seemingly random attitude. Specifically, the mechanism's output is shown to vary with the sensory input only, and not the subject's consent. The subject himself, on the other hand, insists that his beliefs are ultimately the source of the attitude and not a mere variation in the sensory input. To make this more specific, let us look over Caruthers' interpretation of one particularly poignant experiment.

> One robust finding in this literature is that people who have been cleverly manipulated into writing an essay for a paltry sum of money in defense of something that they initially disagree with will end up, after the fact, expressing much more sympathy for the position that they have defended than will other people who were paid a decent amount (Cohen, 1962; Linder et al., 1967) … . Why should writing an essay under conditions of inadequate payment lead to heightened belief, when going through the very same process for adequate pay doesn't? (Note that it can't be the mere fact of thinking up good arguments, and so forth, that produces belief, unless for some reason those who are paid less should argue better!)*(ibid)*

Caruthers explains the above evidence in the following way. The slave subsystems broadcast the following information: this task is requires a lot of effort. I am getting no monetary reward for applying all this effort. The interpreting mechanism would be receiving these broadcasts and applying some theory in order to put out the resultant attitude. For instance, the mechanism might interpret the broadcast in the following way. I don't just apply effort for no reason/reward. My reason must lay in the content of this activity. Hence I really support the idea behind this essay! Importantly, if the slave systems' input is altered, so is the resulting attitude. If the slave systems report that a great monetary reward has been received for this effort expenditure, the interpreting mechanism doesn't need to try to figure out why the effort is being expanded and so doesn't search for reasons within the content of the activity. Hence the attitude "I must really believe in the idea behind this essay" is not produced.

According to Caruthers, the interpretive mechanism would have produced a similar attribution if it were interpreting another person's behavior. For instance if the subject were to see a stranger on the street expanding a great amount of effort writing an essay for little pay, he would have figured that the stranger must really support the idea presented in the essay. Otherwise, why else would he be doing that?

In the next section I will present evidence against ISA. Making self and other attributions of propositional attitudes simply can not be supported by the same kind of mechanism.

Chapter 2. Objections to ISA theory

As we have seen above, Caruthers appeals to empirical findings to defend the Interpretive Sensory Access theory of introspection. His argument takes the form of an inference to the best explanation. In this section I will present empirical findings in neuroscience that show that one of the main predictions of the ISA theory is false. Specifically it cannot be the case that there is a single mental faculty underlying our attributions of propositional attitudes to ourselves and to others. I will show that if the "interpretive mechanism" exists, and moreover that it operates in accordance to Caruthers' description in mind reading tasks, (e.g. detecting external cues and paying attention to others' behavior), then this operation would have to be handled or implemented at the neural level by the Task Oriented Neural Network. On the other hand, it is well known that self-referential thought, including introspective thought is handled by the Default Mode Network. This consequence is problematic for the view that self and other attitude attributions are done by the *same* mechanism. The same cognitive operation can not be implemented by two distinct neural networks that are in competition with one another.

It is important to note that Caruthers would gladly grant that there is an obvious difference between the two tasks (self attributions and other attributions). Self attributions involve making use of apperception and inner speech, while other attribution does not. Could this difference account for the fact that these two tasks are implemented by different neural networks?

Things are not that simple. Remember that, according to Caruthers, a *single* mind reading capacity is used for both self and other attributions. This single capacity relies on

the *same* sub-personal "interpretive" mechanism that takes sensory information as input and produces attitudes as output. To make matters worse, he argues that the only difference between the self attributions and the other attributions is *quantitative* and not *qualitative* in kind. That is, there is simply more sensory evidence available for self attributions than for other attributions. Importantly he insists that the *type* of access to one's attitudes and others' attitudes is the same.

Unfortunately, the Default Mode neural network and the Task Oriented network implement such different types of thinking that they oppose and interrupt one another's functioning. If the only difference between the two networks was that one simply handles a larger *quantity* of information than the other, they wouldn't be in competition. It appears that there is indeed something special about the very nature of self-referential information such that it determines the *type* of operations involved in its processing. In fact, research shows that increased hyperactivity of the DMN network (when the DMN network tries to do the job of the Task Oriented network) is correlated with increased mental dysfunction, progressing from mood disorders to full on schizophrenia. So cases where self and other attribution is really done by the *same* mechanism (and thus implemented by the same neural network) are actually cases of schizophrenia.

Here is a summary of the importance of relevant empirical findings.

a) "Introspective" or rather "self-referential" information is processed by a distinct neural network from all other sensory information. The Default Model Neural network (DMN) handles self-referential information, while all other sensory information (including sensory information about other people) is handled by the Task Oriented network.

b) These two distinct neural networks make possible two distinct types of thinking with two distinct types of phenomenology.

c) The type of thinking made possible by DMN requires different computations to take place than the type of thinking made possible by the task-oriented network.

d) Different computations require different computational mechanisms.

e) The DMN and the task-oriented network compete with one another. Activity in the DMN "interferes" with activity in the task-oriented network by hampering task specific attention and contributing to impaired task performance.

This section will be dedicated to illustrating how "a-e" taken together show the ISA prediction (that there is a single mental faculty underlying our attributions of propositional attitudes to ourselves and to others) to be false. I will start with defining the DMN network, its functional organization, and the cognitive functions it is thought to serve. I will then define the task-oriented network and its relation to the DMN network. Finally, I will introduce what has become known as the "interference" hypothesis. I will elaborate on how fine-tuned and delicate the relation between the DMN and the task-oriented network really is. After explaining how the delicate relationship between these two distinct networks affects normal cognitive functioning, I will introduce an alternative evolutionary explanation for having several *distinct* mindreading mechanisms for self and other attributions.

2.1 INTRODUCING THE DMN

In today's technological age, we might think of ourselves as constantly busy or engaging in some goal-directed activity every second of the day. Even while taking a day off at the beach we are still tapping away on our iphone returning emails or updating our Facebook status. While it is hard to imagine actually being at rest these moments do happen. So what does our brain do when not actively engaged in some goal-directed activity? The waking "resting state" activity of the brain has been termed the "default mode" and is associated with an attenuation of activity in a distinct neural network. Specifically Broyd (et al. 2009) writes that attenuation in the ventral medial prefrontal cortex (MPFC) occurred with tasks involving judgments that were self-referential, while activity in the dorsal medial prefrontal cortex increased for self-referential stimuli, suggesting the dorsal MPFC is associated with introspective oriented thought (Gusnard et al., 2001). Multiple research labs have shown that DMN activity is characterized by very low frequency neuronal oscillations ("low frequency BOLD signal). Moreover, connectivity (which indicates the functional coherency of a neural network) has been show to change with age. For instance there is very limited evidence of DMN activity in an infant brain (Fransson et al., 2007). There is evidence of fragmented connectivity between the DMN regions during rest in young children (7-8 years; Fair et al., 2008) and more consistent DMN connectivity in children aged 9-12 years (Thomason et al., 2008). These neural changes map onto our cognitive development. So while infants have no trouble visually tracking moving objects and the facial expressions of their caretakers, they have trouble tracking their internal mental states.

Fransson (2006) writes that activity in the DMN, in the absence of overt task performance, engages in any number of spontaneous, self-referential mental events such as episodic memory, planning for the future, inner speech, and introspection.

2.2 INTRODUCING THE TASK-ORIENTED NETWORK

The "task oriented" network or the task positive network includes the dorsolateral prefrontal cortex (DLPFC), inferior parietal cortex (IPC) and supplementary motor area (SMA). This network appears to be associated with task-related patterns of increased alertness, response preparation and selection (Fox et al., 2005, 2006a; Fransson 2005, 2006; Sonuga-Barke and Castellanos, 2007). Consider the attentional resources that go into the goal-directed activity of cutting a peanut butter sandwich. Researches have studied vision in the natural world by tracking the subject's gaze (Hayhoe and Rothkopf, 2010). Once the subject is about to cut the sandwich in half, gaze will be directed to the knife handle to guide the hand to pick it up. As the hand closes in on the knife, the eye will move to the corner of the sandwich where the knife tip will be placed to begin cutting. When the subject in engaged in this task, the task-oriented network makes sure that the brain performs the relevant computations that guide gaze allocation and motor movements.

2.3 COMPETITION BETWEEN THE DMN AND THE TASK-ORIENTED NETWORK

Working memory tasks (e.g. solving a puzzle or attentively watching a football game, making a sandwich) that are not self-referential are associated with an attenuation in the DMN regions specified above (Esposito et al., 2006, Salvador et al., 2008).

Should the DMN network not be sufficiently attenuated during this task and a "self-referential" thought surfaces into consciousness (i.e. "I can never manage to make pretty looking sandwich halves" or "am I taking too long with these sandwiches?"), task performance will worsen considerably (i.e. you might end up accidentally cutting your finger instead of the bread).Whenever activity in the task-positive networks of the brain increases (when attention is directed externally), DMN activity has been shown to decrease. Furthermore, spontaneous spikes in the DMN activity (in four left hemisphere regions implicated in self-referential thought) are associated with task unrelated thoughts. These "interferences" are in turn associated with poor task performance (McKiernan et al., 2003, 2006; Corbetta & Shulman, 2006; Fox et al., 2005; Fransson, 2005).

The relationship between the DMN and the task-oriented network is quite fine tuned, and any deviation or imbalance greatly affects normal cognitive functioning of the individual. Broyd cites Sonuga-Barke and Castellanos in stating the following.

> The high degree of temporal anti-correlation emphasizes the potential degree of antagonism between the DMN and the task positive network and the psychological functions they reflect.

The lack of suppression of DMN activity during a goal-oriented task has been associated with various mental disorders. Before we move on to elaborate on the relationship between the two networks another preliminary is in order. It is generally agreed that DMN activity is associated with self- referential thought including mind wandering and

introspection. However, some have proposed that it might also be involved in attributing mental states to others. If DMN were really responsible for both self and other attribution of propositional attitudes, it would appear that there really is no trouble for self/other parity accounts. However, the mental techniques that trigger DMN activity are not those of observing and interpreting others' behavior. It is only when emotionally salient stimuli are involved in simulating others' mental states that DMN shows activation (Maddock, 1999). This sort of self-referential or emotionally salient processing is in fact *attenuated* when attention is directed toward external events (Gusnard and Raichle, 2001).

2.4 DMN DYSFUNCTION IN MENTAL DISORDERS

As illustrated in the previous section, activity in the DMN "interferes" with activity in the task-oriented network. Broyd summarizes the point in the following way: "The ability to maintain attentional focus and resist distraction or lapses of attention is considered to underlie higher order top-down control" (*ibid*). She continues to summarize recent neuroscience findings in the following way: "intrusions of introspective thought produce variability in task performance in normal population" (*ibid*). Groups that suffer the most from most DMN interference are the Schizophrenic patients (Zhou et al., 2007). In general any type of dysfunction of introspective processes is associated with DMN interference. Namely, depressive patients who self-report an abnormally large amount of self-referential thought processing show lack of attenuation of DMN during goal-directed activity. As shown above, this lack of attenuation causes poor task performance. Processing all incoming information in a self-referential manner (e.g. they are all looking

at me, I look silly to them, why do bad things always happen to me) puts a great strain on attentional resources and results in characteristic "absent-minded" behavior exhibited by depressed subjects (Greicious et al., 2007).

As we can see if self and other attributions of propositional attitudes were all handled by the same neural network and consequently the same cognitive mechanism, the world would be filled with "absent-minded" individuals.

2.5 Evolutionary Advantages for Having Distinct Mechanisms for Self and Other Attributions.

Caruthers argues that having the *same* mechanism do the work of self and other attributions is advantageous from an evolutionary perspective. Here is what he writes.

> There exists a good answer to the question why an ''outwardly focused'' mindreading faculty of the sort represented in Figure 1 (or the capacity to construct such a faculty via learning) might have evolved. This is some or other version of the ''Machiavellian intelligence'' hypothesis (Byrne and Whiten, 1988, 1998), which points to the immense fitness advantages that can accrue to effective mindreaders among highly social creatures such as ourselves. We also have good evidence that the brain is constructed in such a way as to realize the global broadcast of perceptual events, thus facilitating other-directed mindreading inter alia, together with introspection of such events as a by-product.

In light of the discussion in the previous sections, I will try to lay out some evolutionary advantages for having distinct mechanisms for handling self and other attributions of propositional attitudes. It appears that not only self and other attributions are not handled by the same mechanism, but that not all other attributions are handled by the same mechanism either. Caruthers' appeal to simplicity in theorizing about a "one for all" mechanism is definitely wrong minded.

Researchers have found that the amount of “self-involvement” (effectively the amount of DMN activated) during a mindreading task depends on our familiarity with the person. For instance, when getting on to the propositional attitude of your best friend you tend to use "Simulation"[2] like techniques which require the use of self-referential oriented DMN network; whereas when getting onto the propositional attitudes of a stranger you tend to use "Theory Theory"[3] like techniques which tax the "task-oriented" neural network. (Krienen et al., 2010). Attributing mental states to individuals perceived to be significant by the attributer relies on systems optimized to process self-referential information. What might be the evolutionary advantage for having specialized mindreading systems instead of a single one?

Immense advantages arise from discerning self-referential information from non-self referential information. It is valuable to know whether to expend valuable energy helping others who are relevant to your survival. For example, monkeys sacrifice themselves by giving off a warning call for other monkeys in their group. By giving off this signal they essential put a big red target on their backs since their cry allows the predator to locate them without fail. This self sacrifice is ultimately advantageous for the individual since most other members of the tribe share a great deal of the genome. Similar sacrifices are often times performed for offspring and so on. Given humans’ richer conceptual repertoire, the classification of “similar” others extends beyond genotype similarity. For instance people who have never met Stalin still sobbed at the news of his illness since they viewed him as a close other. DMN activation during these

[2] This view holds that we represent the mental states and processes of others by mentally simulating them, or generating similar states and processes in ourselves

types of mindreading tasks allows subjects to emotionally identify with others and view others' mental states and relevant to their own.

However, it would not be advantageous to identify or attempt to simulate the mental states of all others. Recall the discussion from the previous sections. If the DMN were activated in all mindreading tasks, the subject's attentional mechanism would greatly suffer. He would no longer be able to perform simple everyday tasks like making a peanut butter sandwich. Moreover, consider the following examples of mindreading.

People who performed atrocities during WW2 had no trouble successfully attributing mental states to their victims for they were able to successfully manipulate and control them. They were able to do so because they felt themselves to be completely "separate" or "dissimilar" to their victims. It is likely that their task-oriented network was active in discerning the mental states of their victims. They were able to methodically observe their victims' behavior and theorize about their attitudes. Once the attitudes of others are well known, they become easier to manipulate.

Given the distinct types of thinking made possible by the DMN and the task oriented network respectively, humans are able to attribute mental states to close others and achieve some degree of shared subjectivity while at the same time accurately attribute mental states and manipulate strangers. It is vital for survival that these two techniques of mindreading remain absolutely separate and do not get mixed up. That is why these two techniques are not both achieved via one and the same mechanism. The swiftness of implementing either technique is so vital (e.g. the self sacrificial act must be

[3] The view that a tacit *theory* (a "folk" psychology) underlies psychological competence

done in a fraction of a second) that a completely distinct neural network is needed for each.

Consider the following remarkable episode documented by a group of researchers studying lions in their natural habitat. A lioness somehow came to "befriend" a gazelle (Animal Planet, 2008). She would defend the gazelle and treat it as a close other e.g. she would groom it and so on. Unfortunately, at the same time the lioness lost the ability to stalk and prey on gazelles. She lost the ability to kill and eventually almost died of starvation. This episode shows the unfortunate reality of the natural habitat in which humans evolved. In this environment, the ability to successfully mindread others and predict their actions and attitudes *without* "identifying" with them is vital for survival. Even within the same species, acts of aggression and calculated manipulation (e.g. during mating) were necessary for the majority of our evolutionary history. In order for each type of attribution to be performed efficiently (whether it be to the self, a close other, or a dissimilar other), two completely separate neural networks had to be developed.

CONCLUSION

The aim of this paper was to show one of the main predictions of the ISA theory to be false. This prediction is the following. There is a single mental faculty underlying our attributions of propositional attitudes, whether to ourselves or to others. I have laid out research in neuroscience that suggests that the above prediction cannot be accommodated. What I did not do is provide a vindication of authoritative and special access. I think that Caruthers has rightly pointed out that the transparency model is lacking in accounting for the prevalence of confabulation in *normal* subjects. Moreover, he is right to demand an empirical basis for the so called non-extravagant mechanism for acquiring self-knowledge. I think that there is indeed something special or rather unique about the processing mechanism responsible for handling self-relevant information. The very cognitive process by which this information is handled is distinct from that responsible for tracking external features in the environment. These findings however do not show whether the uniqueness of this mechanism is good or bad for the security status of self-knowledge. All they show is that Caruthers' broadcasting model is on the wrong track.

I have also presented an alternative evolutionary explanation that outlines the adaptive advantage for having at least two attributive mechanisms. Caruthers may have been overly motivated by research in Artificial Intelligence and may have forgotten to consult our neurological constraints.

REFERENCES

Andrews-Hanna, J. R., The Brain's Default Network and Its Adaptive Role in Internal Mentation, 2011, Neuroscientist 2012 18:251

Baars, B., Ramsoy, T. and Laureys, S., Brain, consciousness, and the observing self, Trends in Neurosciences 2003 26: 671–675.

Broyd, J. S., Demanuale, C., Debener, S., Default-mode Brain Dysfunction in Mental Disorders: A Systematic Review, Neuroscience and Behavioral Reviews 2009 33:279-296

Carruthers, P., The Opacity of Mind: An Integrative Theory of Self-Knowledge, Oxford University Press 2011

Carruthers, P., Introspection: Divided and Partly Eliminated, Philosophy and Phenomenological Research 2010 80(1):76-111

Castellanos, F.X., Margulies, D.S., Kelly, A.M.C., Uddin, L.Q., Ghaffari, M., Kirsch, A., Shaw, D., Shehzad, Z., Di Martino, A., Biswal, B.B., Sonuga-Barke, E.J.S., Rotrosen, J., Adler, L.A., Milham, M.P., Cingulate-precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder. Biol. Psychiatry 2008 63:332–337.

Castellanos, F.X., Sonuga-Barke, E.J.S., Scheres, A., Di Martino, A., Hyde, C., Walters, J.R., Varieties of attention-deficit/hyperactivity disorder-related intraindividual variability. Biol. Psychiatry 2005 57:1416–1423.

Esposito, F., Bertolino, A., Scarabino, T., Latorre, V., Blasi, G., Popolizio, T., Tedeschi, G., Cirillo, S., Goebel, R., Di Salle, F., Independent component model of the default-mode brain function: assessing the impact of active thinking. Brain Res. Bull. 2006 70:263–269.

Fair, D.A., Cohen, A.L., Dosenbach, N.U.F., Church, J.A., Miezin, F.M., Barch, D.M., Raichle, M.E., Petersen, S.E., Schlaggar, B.L., The maturing architecture of the brain's default network. Proc. Natl. Acad. Sci. U.S.A. 2008 105:1028–1032.

Fox, M.D., Raichle, M.E., Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging, Nat. Rev. Neurosci. 2007 8:700–711.

Fox, M.D., Snyder, A.Z., Raichle, M.E., Global signal regression and anticorrelations in resting state fMRI data, In: Poster Presented at Human Brain Mapping Conference June 2008, Melbourne, Australia.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005., The human brain is intrinsically organised into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. U.S.A. 102, 9673–9678.

Fox, M.D., Snyder, A.Z., Zacks, J.M., Raichle, M.E., Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses, Nat. Neuroscience 2006 9:23–25.

Fox, M.D., Corbetta, M., Snyder, A.Z., Vincent, J.L., Raichle, M.E., Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. Proc. Natl. Acad. Sci. U.S.A. 2006 103:10046–10051.

Fransson, P., Spontaneous low-frequency BOLD signal fluctuations: an fMRI investigation of the resting-state default mode of brain function hypothesis, Human Brain Mapping 2006 26:15–29.

Fransson, P., How Default is the Default Mode of Brain Function? Further Evidence From Intrinsic BOLD Signal Fluctuations, NeuroPsychologia 2006 06:117

Greicius, M.D., Flores, B.H., Menon, V., Glover, G.H., Solvason, H.B., Kenna, H., Reiss, A.L., Schatzberg, A.F., Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulated cortex and thalamus, Biological Psychiatry 2007 62:429–437.

Greicius, M.D., Kiviniemi, V., Tervonen, O., Vainionpa¨a¨ , V., Alahuhta, S., Reiss, A.L., Menon, V., Persistent default-mode network connectivity during light sedation. Human Brain Mapping 2008 29:839–847.

Greicius, M.D., Supekar, K., Menon, V., Dougherty, R.F., Resting-state functional connectivity reflects structural connectivity in the default-mode network. Cerebral Cortex 2009 19(1):72–78.

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proceedings of the National Academy of Science U.S.A. 2008 100:253–258.

Greicius, M.D., Menon, V., Default-mode activity during a passive sensory task: uncoupled from deactivation but impacting activation, Journal of Cognitive Neuroscience 2004 16:1484–1492.

Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V., Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI, Proceedings of the National Science Academy U.S.A. 2004 101:4637–4642.

Gusnard, D.A., Akbudak, E., Shulman, G.L., Raichle, M.E., Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function, Proceedings of the National Science Academy U.S.A. 2001 98:4259–4264.

Gusnard, D.A., Raichle, M.E., Searching for a baseline: functional neuroimaging and the resting human brain, National Review of Neuroscience 2001 2:685–694

Hayhoe, M., Rothkopf, C., Vision in the Natural World, Cognitive Science 2010 2(2)

Krienan, F. M., Tu, P., Buckner, R., Clan Mentality: Evidence That the Medial Prefrontal Cortex Response to Close Others, The Journal of Neuroscience 2010 30(41):13906-12915

Lieberman, M. D., Social Cognitive Neuroscience: A Review of Core Processes, Annual Review of Psychology 2007 58:259-289

Lycan, W., Consciousness, MIT Press 1987

Animal Planet, Lionness Adopts Gazelle, Mutual of Omaha, Discovery Channel 2008

Maddock, R.J., The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain, Trends Neuroscience 1999 22:310–316.

McKiernan, K.A., D'Angelo, B.R., Kaufman, J.N., Binder, J.R., 2006. Interrupting the 'stream of consciousness': an fMRI investigation. NeuroImage 29, 1185–1191.

Salvador, R., Martı´nez, A., Pomarol-Clotet, E., Gomar, J., Vila, F., Sarro´ , S., Capdevila, A., Bullmore, E., A simple view of the brain through a frequency-specific functional connectivity measure, NeuroImage 2008 39:279–289.

Saxe, R., Against Simulation: The Argument From Error, TRENDS in Cognitive Science 2006 9(4):174-179

Schacter, D., Adaptive Constructive Processes and the Future of Memory, American Psychologist, in press

Sonuga-Barke, E.J.S., Castellanos, F.X., 2007. Spontaneous attentional fluctuations in impaired states and pathological conditions: a neurobiological hypothesis. Neurosci. Biobehav. Rev. 31, 977–986.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Default Network Activity Coupled with the Frontoparietal Control Network, Supports Goal-directed Cognition, NeuroImage 2010 53:303-317

Schwitzgebel, E., Introspection, The Stanford Encyclopedia of Philosophy 2010